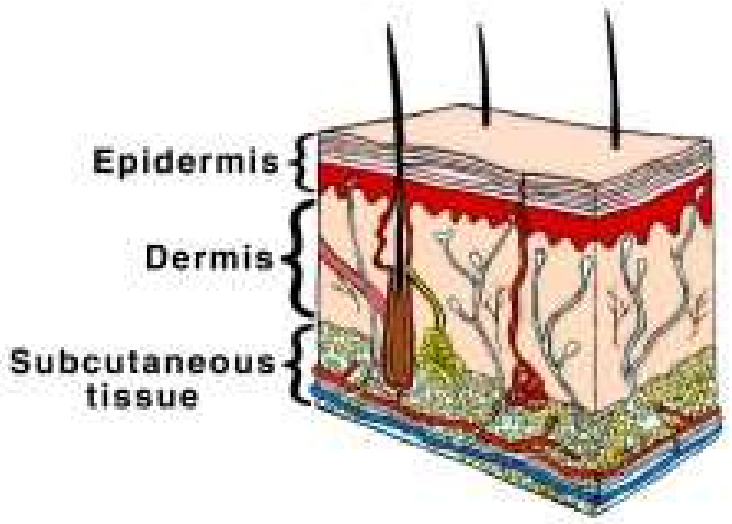
Activation Functions

Deep Learning

Magesh anand D

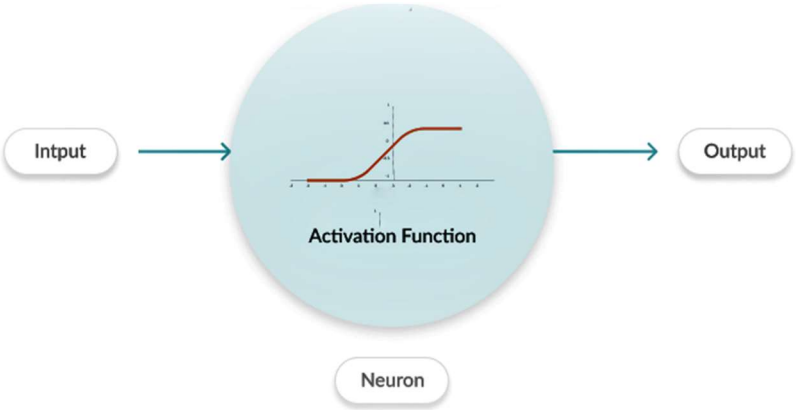
Neural network are made up of artificial neuron similar in concept to neuron in human body. These neuron are connected to each other, forming the lot of connection and the system works through the activation of neurons.

Human skin composed of several layer. Very top layer is the **epidermis**. The epidermis contains very sensitive cell called touch receptors that give the brain a variety of information about the environment the body is in.



And is controlled by the **somatosensory** system. This system is responsible for all the sensations we feel- cold, hot smooth, rough, pressure, itch, pain, vibrations, and more. If human body was get into the hot environment then hot sense neuron will get activated give an information to brain **HOT** environment. If is cold then sense neuron will get activated give an information is **COLD**.

Likewise artificial neural network work base weights. Neuron will get activated if is high weights it won't get activated if low weight. Neuron activation base on the some **ACTIVATION FUNCTION** have to apply into the neural network. Activation function is used to propagate output from one layer to next layer including the output layers. There are many different activation function base on problem the activation function will get changed in neural network.

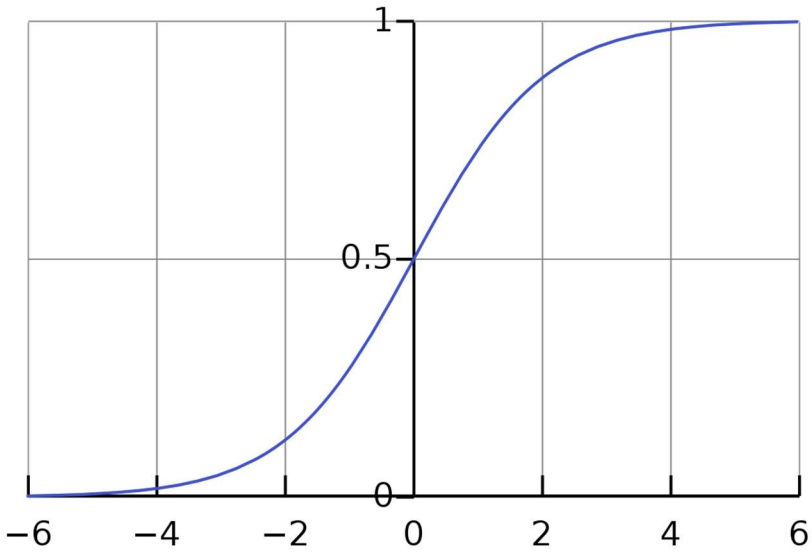


List of activation function:

1. Sigmoid function
2. TanH function (**Hyperbolic Tangent**)
3. ReLU function (**Rectified linear unit**)
4. Leaky ReLU function
5. ELU function
6. PReLU function
7. RReLU function
8. Softmax function
9. Maxout function
10. Softplus function
11. Swish function
12. Mish function

Sigmoid Function:

The sigmoid activation function also known as called Logistic function, is most famous non- linear activation function for neural network. The input to the function is transformed into the value between 0 to 1. But it's not a probability function but its give values based on probability of data. In simple term if the values is greater than 0.5 output is 1 & if less than 0.5 output is 0.



Formula: f(x) = 1/ (1+e^-x)

Advantages of Sigmoid function:

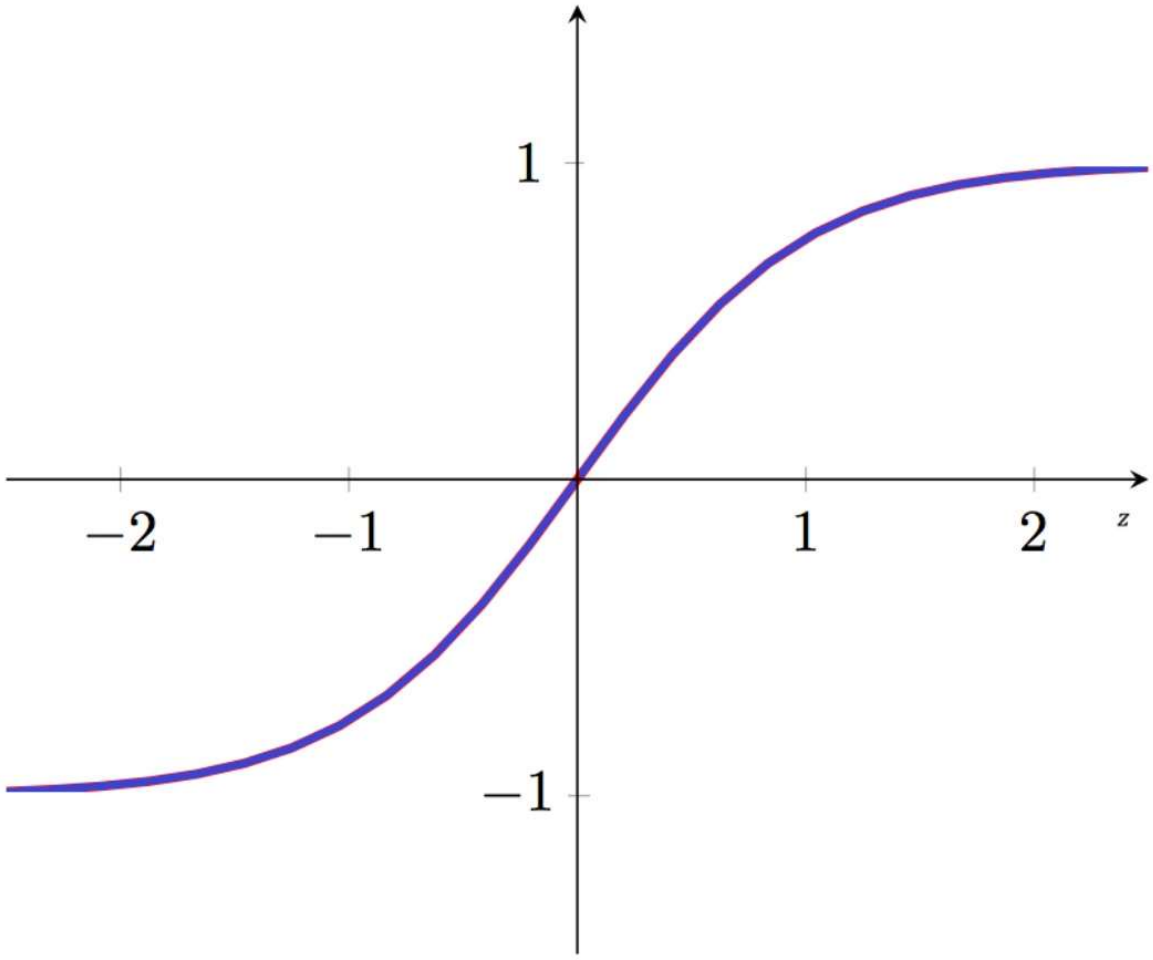
- Sigmoid function can reduce the extreme values or outliers in the data without removing them all.
- Sigmoid function is convert independent variable of near infinite range into simple probabilities between 0 and 1.
- This function is mostly used in terms of binary class classification because of the output fall under the 0 and 1. But it fail in multi class classification

Disadvantage of Sigmoid function:

- This function work based on exponential operation. It give the smooth curve. But for very high or very low values of **X** there is almost no change to the prediction causing the **Vanishing gradient** problem.
- During the back propagation is very difficult to update the weights while taking derivative of features with help of loss function.
- Output is not zero centered which mean the values will not fall under the zero.

TanH / Hyperbolic Tangent function:

The TanH function is very similar function to sigmoid function. The value fall under the -1 to +1. Unlike the Sigmoid function the normalized the data which mean reduce the extreme values or outliers in the data without removing them all but in between values -1 to 1.



Formula f(x) = (e^x – e^x) / (e^x + e^x)

Advantage of TanH function:

- TanH function is convert independent variable of near infinite range into simple probabilities between **-1 to +1.**

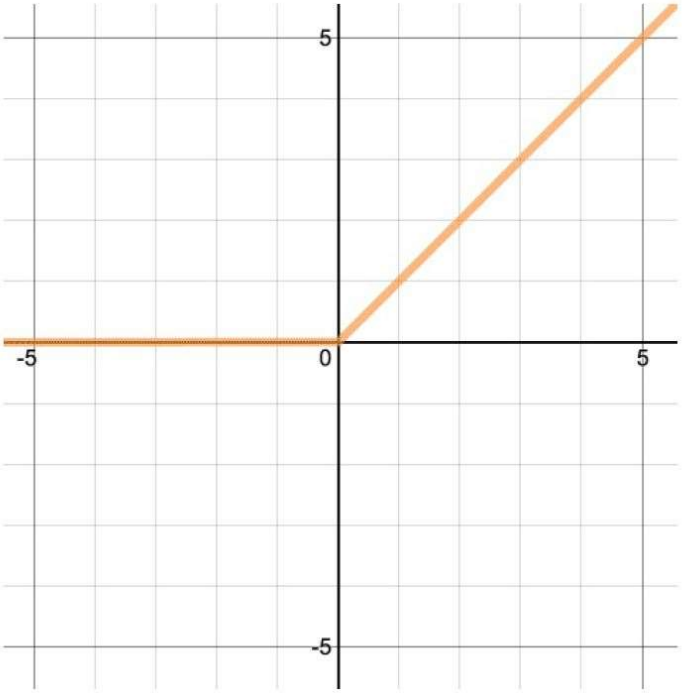
- The tanh is that it can deal more easily with **negative** numbers.
- The output is the zero centered which mean the function output values that have strongly negative **(-1)**, neutral **(0)** and positive **(+1).**
- We can also use this function into binary class classification function like sigmoid function.

Disadvantage of TanH function:

- Same like sigmoid this function work based on exponential operation. It give the smooth curve. But for very high or very low values of **X** there is almost no change to the prediction causing the **Vanishing gradient** problem.
- During the back propagation is very difficult to update the weights while taking derivative of features with help of loss function.

ReLU (Rectified linear unit):

Rectified linear is more interesting transform that activates a neuron only if the input is above a certain quantity. If the input of the feature is below zero the function is consider has **zero**. ReLU is **non-linear** function and has the advantage of not having any backpropagation error like sigmoid.



Formula: F(x) = max (0, 1)

Advantage of ReLU function:

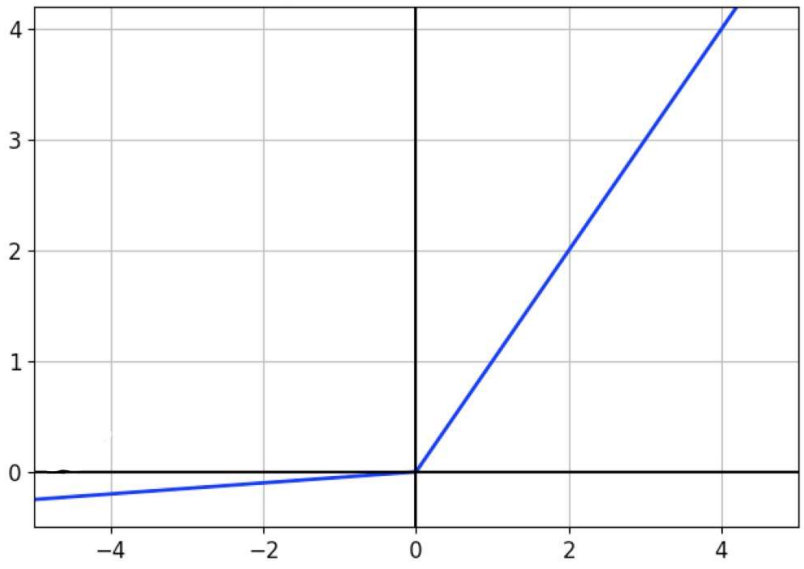
- If the input is positive vanishing gradient problem wont occurs in ReLU function. Because there is always slope in function help to find the derivatives and change the **weights** during the **backpropagation**.
- The Computational speed is much faster compare to sigmoid and TanH function.

Disadvantage of ReLU function:

- When there is negative input the ReLU activation function constant in zero. During the backpropagation the vanishing gradient problem will occur.
- There is no slope in case of negative input we won't find the derivatives to update the weights during the backpropagation.
- ReLU is not a zero centric function like TanH.

Leaky ReLU:

Leaky **ReLU** are strategy to mitigate the dying ReLU issue. As opposed to having the function being the zero when X < 0, the leaky ReLU will instead have the small negative slope around **(0.01).**



Formula: f(x) = max (0.1*x, x)

Advantage of Leaky ReLU Function:

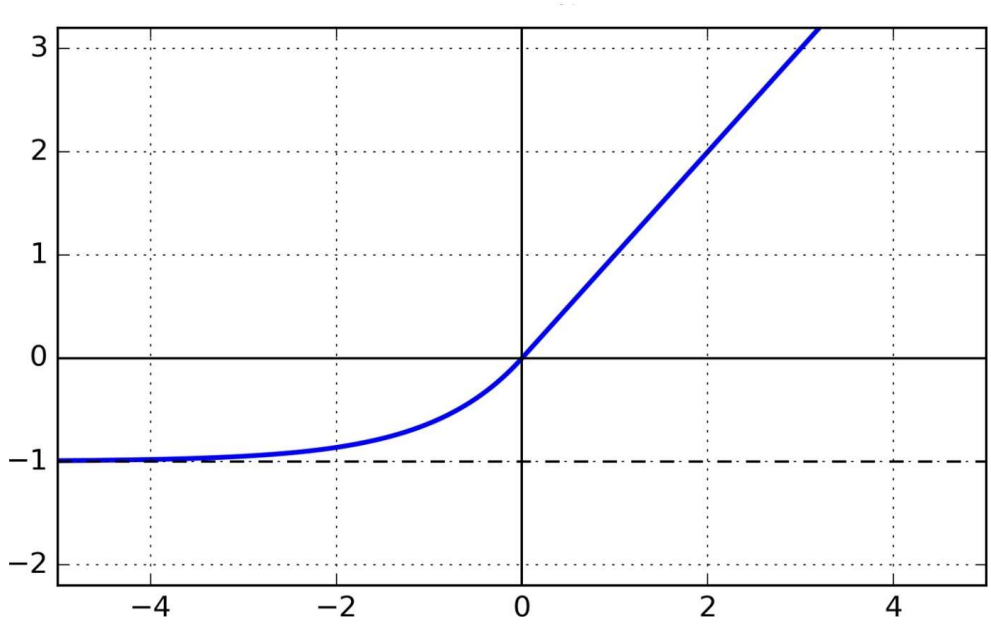
- Prevents dying ReLU problem which mean there is small positive slope in negative area.
- So its dose enable backpropagation even for the **negative** input values.
- Otherwise is ReLU activation function

Disadvantage of Leaky ReLU function:

- Result not consistent Leaky ReLU does not provide consistent prediction for the negative input values.
- There is only small slope during the backpropagation for negative input not much derivatives while updating the weights.

ELU (Exponential Linear units) function:

ELU is one of the activation function are used to solve the dying ReLU issue. ELU is consists of two different equations. At the same time two separate derivatives should be calculated for ELU activation function.



Formula: f(x) = X if X ≥ 0, f (x) = α (e^x-1)

Advantage of ELU Activation function:

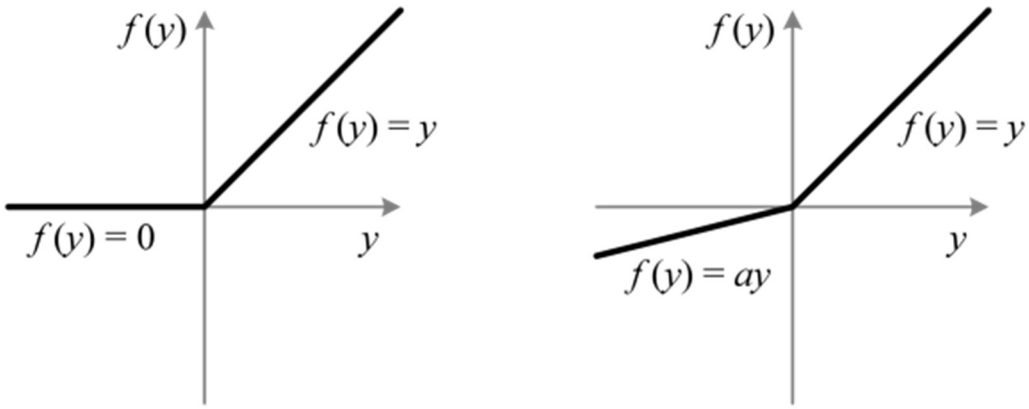
- Prevents dying ReLU problem.
- No vanishing gradient problem.
- Mean of the output is almost close to zero. This is one of the best function for normalization the data.
- ELU is zero centric function.
- There is always some kind of slope in terms of both positive and negative values help to find the derivatives during the backpropagation.

Disadvantage of ELU Activation function:

- Same likewise Leaky ReLU saturates for the large negative value. Which mean small problem is that it is slightly more computationally intensive in negative values.
- In theoretical ELU is good. There is no physical evidence ELU is good function better than ReLU.

PReLU (Parameters Rectified linear unit) activation function:

PReLU is work base on the Leaky ReLU activation function and there is small changes towards the negative inputs. Instead of multiplying **0.01** values in negative inputs the negative part adaptively learn themselves by alpha **(α)** in input.



Formula: $f(y_i) = (y_i \text{ if } y_i > 0 \ \& \ \alpha_i y_i \text{ if } y_i \leq 0)$

Advantage of PReLU Activation function:

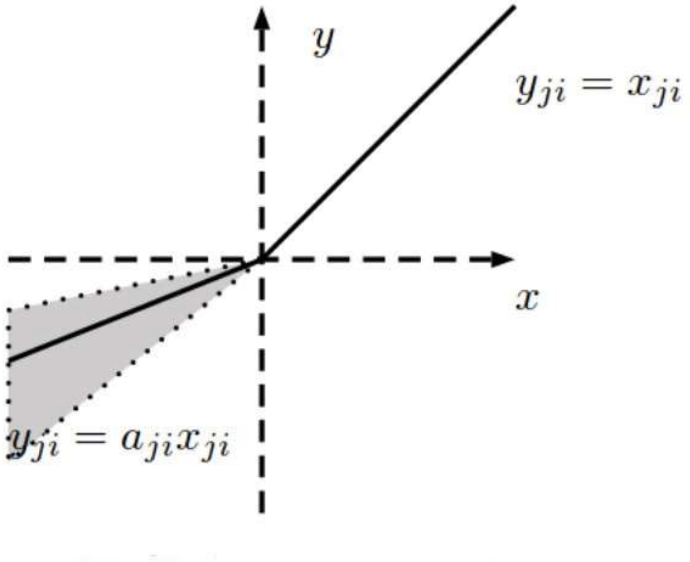
- If $\alpha_i = 0$ f is become ReLU and if $\alpha_i > 0$ f become leaky ReLU.
- The negative slope is α_i learnable parameter
- There is no vanishing gradient problem

Disadvantage of PReLU Activation function:

- Same likewise Leaky ReLU saturates for the large negative value. Which mean small problem is that it is slightly more computationally intensive in negative values even though there is learnable parameters in activation function.

RReLU (Randomized ReLU) Activation function:

Randomized rectified linear unit activation function is same like the PReLU activation function. In PReLU alpha **(α)** consisted as learning parameter based on the input features but in RReLU **(α)** is randomized input values bases on the **uniform** distribution.



Formula: $f(x) = (x \text{ if } x > 0 \ \& \ \alpha x \text{ if } x \leq 0)$

Advantage of RReLU Activation function:

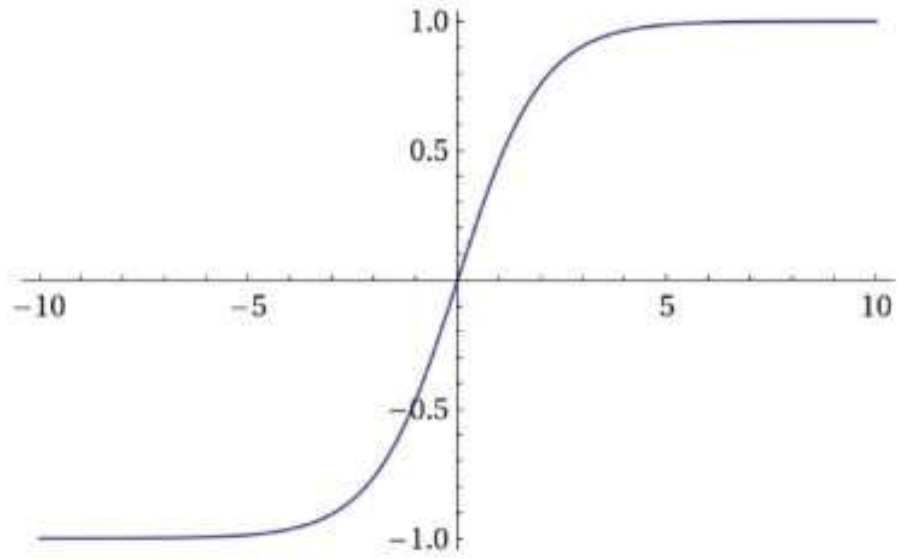
- Negative slope is **randomized** learnable parameter.
- There is always some kind of slope for taking the derivatives during the backpropagation.
- This is zero centric activation function.

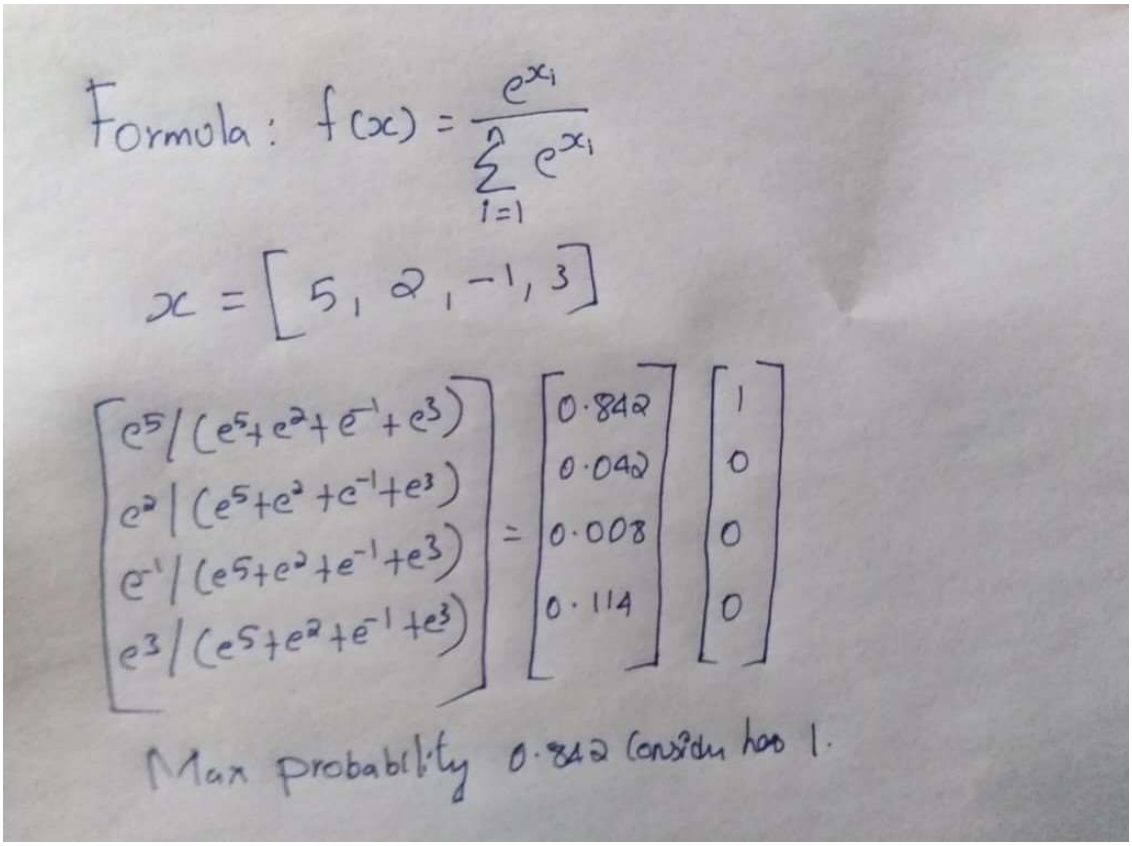
Disadvantage of RReLU activation function:

- Same likewise PReLU saturates for the large negative value. Which mean small problem is that it is slightly more computationally intensive in negative values even though there is randomized learnable parameters in activation function.

Softmax Activation function:

Softmax is a generalization of logistic regression inasmuch as it can be applied to continuous data (rather than classifying binary) and can contain multiple decision boundaries. This function work based on the probability of the input features. It handles multinomial labelling systems. Softmax is the function you will often find at the output layer of a classifier.



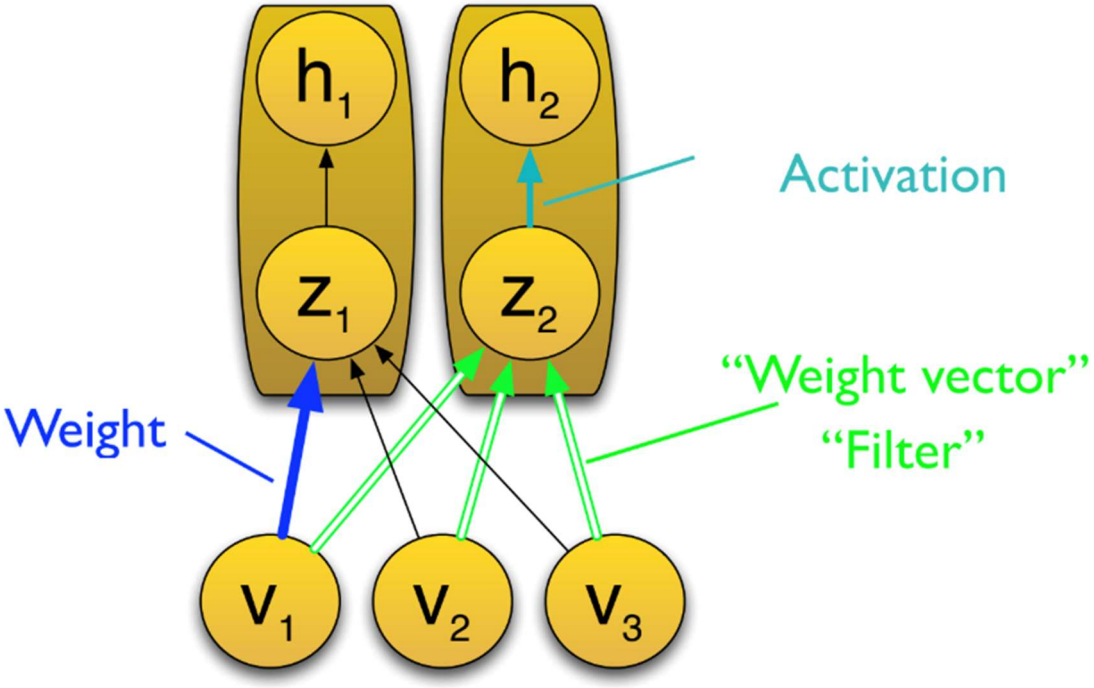


Advantage of Softmax Activation function:

- Softmax function combines the all factors of original values and outputs gives as probabilities values
- Most common activation for the multiclass classification activation function based on probabilities values.

Maxout Activation function:

Maxout activation function is replace traditional activation function such as Sigmoid, tanh or ReLU. Which mean those function are generalized activation like pre-determined function but maxout is truly based on weights of input features. Its gives the max of positive or negative values after the train the input features based on bias and weights.



Formula: f(x) = max ($W_i X_i + b$)

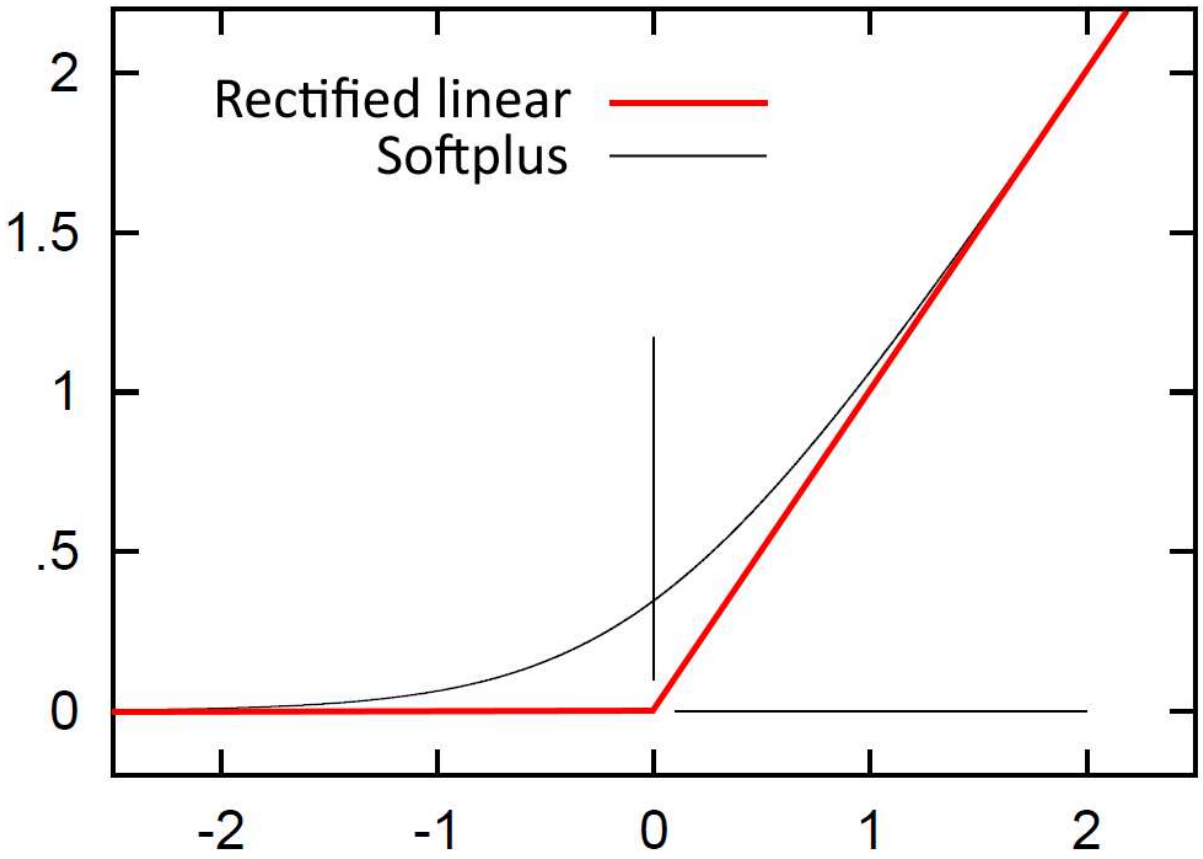
Advantage of Maxout activation function:

- Maxout activation is combination of both **ReLU** and **Leaky ReLU** activation function.
- Its gives Max positive values like ReLU and Max negative value like Leaky ReLU after the train the input features based on weight and bias.
- There is no vanishing gradient problem.

Softplus activation function:

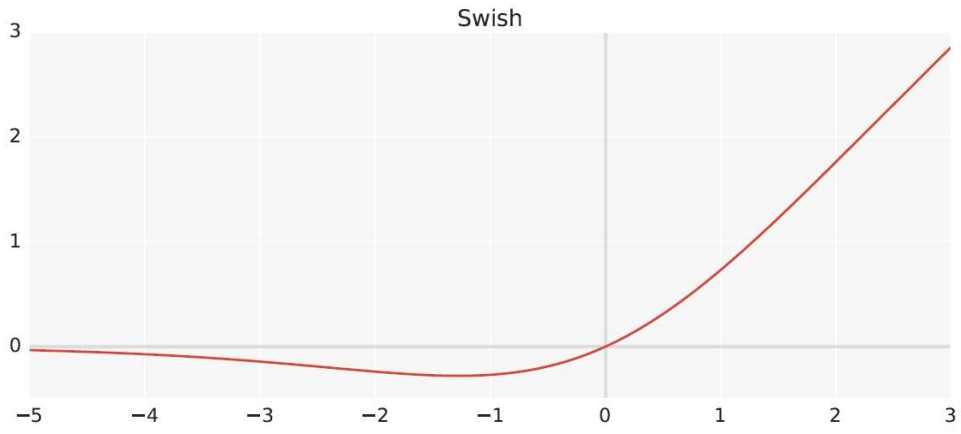
This activation function is considered to be the smooth version of ReLU activation function. It is unilateral suppression like ReLU. It has wide acceptance range (0,+ ∞). Softplus activation function have similar kind of feature like ReLU but its smooth curve.

Formula: f(x) = log (1+e^x)



Swish (A Self-Gated) Activation function:

Currently the most successful and widely used function is ReLU activation function. **Google Brain** team has proposed a new activation function is swish activation function. It's inspired by **sigmoid** function. Their experiment shows the swish activation function work better than ReLU on deep models with datasets. Best use case **ImageNet**, Mobile **NASNet**, **LSTM**.



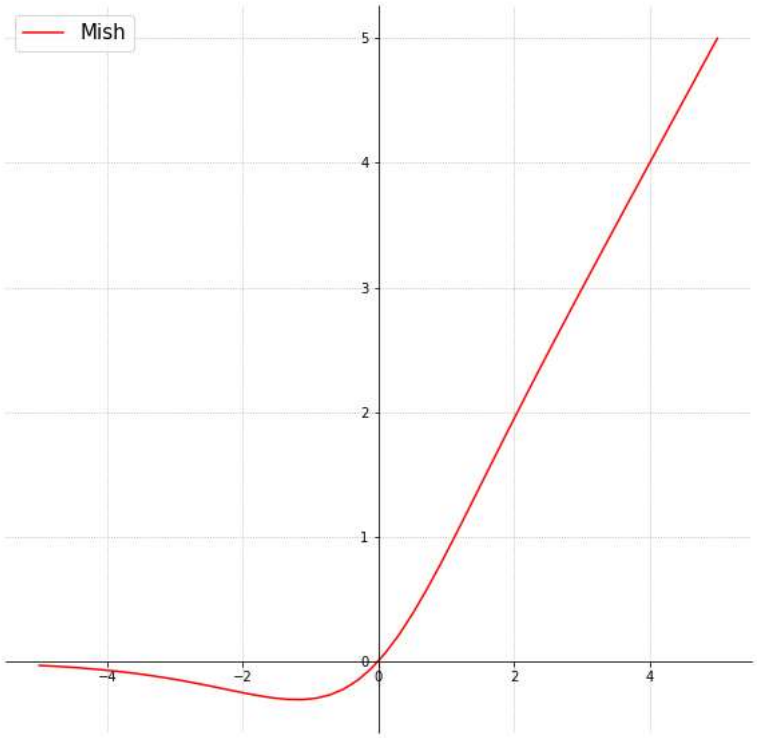
Formula: f(x) = x.sigmoid(x)

Advantages of Swish Activation function:

- There is no vanishing gradient problem
- There is always slope even in negative values its help to find the derivatives during the backpropagation.
- Swish curve is always be smooth and non-monotonic. The non-monotonic property is make different from other common activation function and there is no stripes line in the function.

Mish activation function:

Mish is a self-Regularized Non-Monotonic neural activation function. Mish is one of the most recent activation function. This function is combination of tanh and softplus activation function. Most experiment suggest the mish work better then **ReLU**, **sigmoid** and better then **swish** activation function.



Formula: f(x) = x*tanh (softplus(x))

Advantages of Mish activation function:

- There is no vanishing gradient issues.
- There is always slope even in negative values its help to find the derivatives during the backpropagation
- There is strong regularization effect and reduce over fitting model.